A COMPARISON OF SCATTER-DIAGRAM ANALYSIS WITH DISCRIMINANT ANALYSIS AND A NOTE ON MAXIMIZING THE SKILL SCORE

HARRY R. GLAHN and ROGER A. ALLEN
U.S. Weather Bureau, Washington, D.C.

ABSTRACT

Two questions concerning statistical objective forecasting systems are investigated: (1) Does multiple discriminant analysis (MDA) yield better results than do scatter-diagram analyses made by meteorologists experienced in their use? (2) How should categorical forecasts be derived from probability estimates when it is desired to maximize the Heidke skill score? It is determined that for the particular case tested, a small sample of data and two predictor variables, a subjective analysis will usually yield a better probability forecast than MDA. A method devised by Bryan for making categorical forecasts from probability estimates appears to give better results on test data than do two other methods investigated.

1. INTRODUCTION

In this paper two questions concerning statistical objective forecasting systems are investigated: (1) Does multiple discriminant analysis (MDA) yield better results than do scatter-diagram analyses made by meteorologists experienced in their use? (2) How should categorical forecasts be derived from probability estimates when it is desired to maximize the Heidke skill score? Data samples compiled by Williams [7] were used for empirical testing. These data were vorticity and vorticity advection values at 500 mb. at Salt Lake City and the occurrence or nonoccurrence of precipitation at Salt Lake City during the next 12-hr. period.

2. MDA AND SCATTER-DIAGRAM ANALYSES

For many years studies have been conducted which attempt to find statistical relationships of weather variables to other atmospheric data. Such studies have been in the areas of (conditional) climatology, synoptic climatology, and objective forecasting techniques. One method of statistical data analysis which has been very useful for producing objective forecasting procedures is the scatter-diagram. In recent years, particularly since the advent of the electronic computer, more objective methods of analysis such as multiple regression and discriminant analysis have been used extensively. If the probability of occurrence of a categorical variable is desired, scatter-diagram and discriminant analysis offer alternative methods of solution. There seem to have been few direct comparisons of these two techniques with the use of real meteorological data.

Discriminant analysis for multiple groups as developed by Bryan [3] and Rao [6] and applied by Miller [5] has the advantage of being able to handle many predictors as easily as a few, provided the computer time is available. A single scatter-diagram can accommodate either one or two predictors, and if more predictors are to be used combinations of scatter-diagrams are needed [1]. The discriminant analysis produces one or more discriminant functions; at this point a major problem is yet to be solved—the estimation of predictand probabilities with the use of these functions. Miller [5] has used the procedure of plotting all sample points in the discriminant space and using as the probability estimate for each predictand group the relative frequency of that group in the region about the point defined by the predictors. The region about the point is defined as that containing the closest k sample points in terms of Euclidean distance. k is large enough to provide a stable estimate but small in comparison to the total sample. The use of the distance concept is quite time-consuming, but, because of the binary nature of computers, the computer time required is greatly reduced if the predictors are all binary or dummy variables. In the process of coding a continuous predictor into several binary predictors usually some of the information in the original predictor is discarded but, at the same time a degree of nonlinearity in the predictand-predictor relationship is allowed.

Discriminant analysis was performed on a 396-case sample of data which was essentially that described by Williams [7]. The vorticity and vorticity advection variables were coded into cumulative binary predictors as indicated in tables 1 and 2.

The three predictors selected by screening were (1) vorticity No. 4, (2) vorticity advection No. 4, and (3) vorticity No. 3, in that order. The cutoff procedure described by Miller [5] with $\alpha=0.05$ was used. These predictors divide the predictor space into six regions as shown in figure 1; each binary predictor divides the space

TABLE 1.—*Binary coding for vorticity*

Vorticity (rounded to units $\times 10^{-5}$)	Binary Predictor No.					
	1	2	3	4	5	6
≥16	1	1	1	1	1	1
14, 15	0	1	1	1	1	
12, 13	0	0	1	1	1	
10, 11	0	0	0	1	1	
8, 9	0	0	0	0	1	
6, 7	0	0	0	0	0	
≤5	0	0	0	0	0	

TABLE 2.—*Binary coding for vorticity advection*

Vorticity advection (rounded to units $\times 10^{-5}$)	Binary Predictor No.				
	1	2	3	4	5
≥5	1	1	1	1	1
3, 4	0	1	1	1	1
1, 2	0	0	1	1	1
0	0	0	0	1	1
−1, −2	0	0	0	0	1
≤−3	0	0	0	0	0

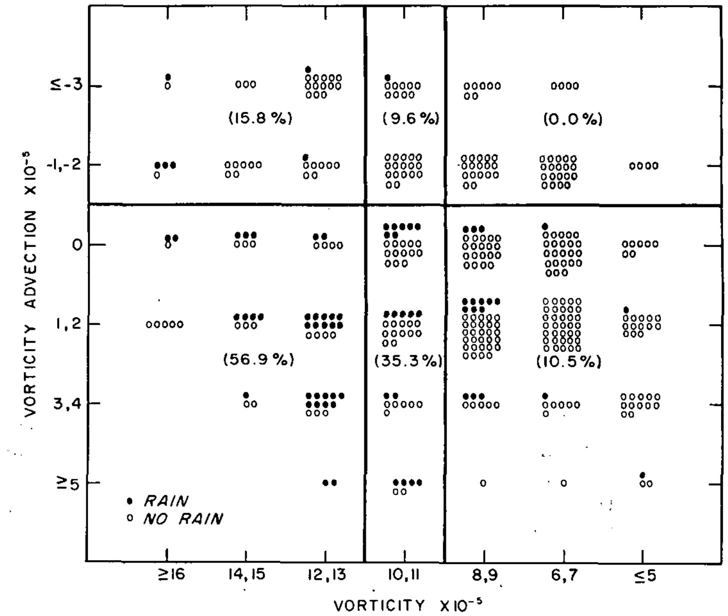

FIGURE 1.—Scatter-diagram of the development data. The six regions defined by MDA are indicated by the vertical and horizontal lines. The estimate of the probability of rain given by MDA for each of the six regions is shown.

with a vertical or a horizontal line. The probability estimates given by MDA are indicated within each region.

Since it was desired to compare the MDA analysis with subjectively made analyses, 10 meteorologists were asked to make analyses of the raw data as plotted in figure 1; no information was furnished other than a copy of this plotted chart without the MDA-derived information. Each analyst had had experience with scatter-diagrams and each was asked to provide an estimate of the conditional probability of rain at each of the 42 points in the predictor space. One of the 10 subjective analyses is shown in figure 2.

Prior to the MDA and subjective analyses it was decided to use the *P*-score [2] as the basis for comparison on a 482-case test sample of data also compiled by Williams [7], and to make two statistical tests. In each test a null hypothesis was to be tested against an alternative:

(1) H_0: The probability of a subjective analysis producing a better score than MDA on the test sample is ½.
  H_1: All alternatives.

(2) H_0: The mean of the subjective scores is the same as that of MDA on the test sample.
  H_1: All alternatives.

If the possibility of *exactly* the same score occurring for a subjective analysis as for MDA is ruled out, the binomial distribution can be used as the basis for a two-tailed significance test for (1). If it is assumed that the subjective scores on test data are distributed as $N(\mu_2, \sigma^2)$, the two-tailed *t*-test can be used to test $H_0{:}\mu_1{=}\mu_2$ against $H_1{:}\mu_1 \neq \mu_2$ where μ_1 is the score made by MDA on the test sample.

In the significance testing it is not assumed that the test sample is a random sample from some population nor are conclusions drawn about that population. The development and test samples are fixed and since MDA (with its required inputs which are considered here as fixed) is completely objective, its score on the test sample is a constant. The sampling, which allows and necessitates statistical testing, is that of sampling randomly the scores, computed on test data, of objective forecasting systems defined by analyses made by meteorologists experienced in the art of scatter-diagram analysis. The conclusions, therefore, apply to only these two samples of data, one for development and one for test, and to the population of meteorologists defined above.

There are two assumptions underlying each statistical test: (1) the meteorologists were randomly chosen from the population as defined, and (2) the form of the distribution is binomial in one case and normal in the other. With proper sampling it seems the binomial is the correct distribution to use in the former case; however, in the latter case the scores may not be normally distributed.

The *P*-scores computed on test data from the scatter-diagram analyses and MDA are listed in table 3. Only one *P*-score for the subjective analyses was worse than that for MDA. This result gives an alpha region (significance level) of size 0.021 for the two-tailed binomial test. However, the *t*-test failed to show significance at the 10 percent level. The probable reason for this result was the somewhat rectangular rather than normal distribution of the sample *P*-scores; consequently the estimated variance in the denominator of *t* was larger than it would have been had the observed *P*-scores been more nearly normally distributed. This "apparently" non-normal distribution does not negate the results of

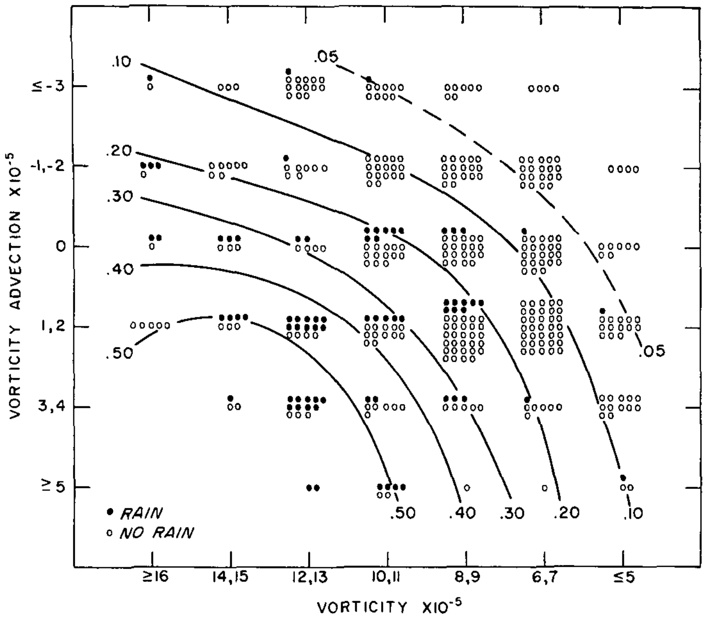

FIGURE 2.—One of the 10 subjective analyses of the data plotted in figure 1.

the test but rather tends to explain the seemingly contradictory results of the two tests.

3. DETERMINATION OF CATEGORICAL FORECASTS FROM PROBABILITY DISTRIBUTIONS

Three methods of determining a categorical forecast of precipitation from the 10 scatter-diagram and MDA probability distributions were investigated. The first was one many times used by forecasters: if the probability of rain equals or exceeds 50 percent, rain is forecast. The second method was that of determining the breakpoint between rain and no-rain forecasts which maximizes the skill score on dependent data. The third method was that devised by Bryan [4]. This last method assumes that the probability estimates are unbiased and a breakpoint can be found for the two-category case that purports to maximize the skill score on independent data. The skill scores for the 11 probability-estimating procedures and the three probability-to-categorical transformations are shown in table 4.

TABLE 3.—*P-scores for the 10 subjective analyses and MDA*

Forecast procedure	*P*-score
Subjective Analysis:	
No. 1	0.2042
No. 2	.2068
No. 3	.2072
No. 4	.2090
No. 5	.2097
No. 6	.2098
No. 7	.2117
No. 8	.2119
No. 9	.2141
No. 10	.2149
MDA	.2148

TABLE 4.—*Test data skill scores for 3 methods of probability-to-categorical forecast transformations*

Forecast procedure	Skill scores		
	0.5 breakpoint	Best breakpoint for dependent data	Bryan method
Subjective analysis:			
No. 1	0.3794	0.3794	0.4122
No. 2	.3979	.3405	.3963
No. 3	.3820	.3820	.4038
No. 4	.2506	.4039	.3684
No. 5	.2967	.3820	.4065
No. 6	.3338	.3338	.4027
No. 7	.2342	.3557	.3653
No. 8	.3684	.3075	.4065
No. 9	.3170	.4133	.4133
No. 10	.3338	.2723	.3526
MDA	.3301	.3929	.3929

A paired t-test was performed for each pair of columns of skill scores in table 4. In each case the null hypothesis H_0: $\mu_1-\mu_2=0$ was tested against all alternatives H_1: $\mu_1-\mu_2\neq 0$ where μ_1 and μ_2 are the means of populations from which the scores in the respective columns were taken. For the test to be valid the population paired differences must be normally distributed. The t-values indicated the breakpoint computed by Bryan's method is better than either the 50 percent breakpoint or the best breakpoint for the dependent data, the significance levels being about 0.1 and 2 percent respectively. Although the 50 percent breakpoint was not as good for this sample as that determined to be the best on the dependent data, the difference was significant at only the 20 percent level.

4. CONCLUSIONS

The case studied here was rather simple in that the data sample was small, only two continuous predictors were involved, the predictand was dichotomous, and therefore the scatter-diagram analysis was relatively uncomplicated; these conditions must be considered when drawing conclusions from the results presented here. However, this case is typical of many in which objective forecasting procedures are required, but, because of the limitations of data availability or data-collecting capabilities, only small samples can be used and only a few simple predictors are to be tested.

The indications are that, for the specific samples used, meteorologists familiar with scatter-diagram analysis can be expected to devise a better probability forecasting method, judged by the P-score, than can be devised with binary-predictor MDA. Binary-predictor MDA can divide the predictor space only by specifying "vertical" and "horizontal" lines, while subjective analyses can introduce curved lines or straight diagonal lines. This allows the probability estimate at a particular point to be influenced by observations in a region which may be elongated in a diagonal direction, or irregular in some way determined by the judgment of the analyst. All ten subjective analyses used diagonal lines roughly similar to those of figure 2.

For these samples of data, the method devised by

Bryan for transforming probability estimates into categorical forecasts to maximize the Heidke Skill Score is superior to the use of either the 50 percent breakpoint or the best breakpoint on the development data sample. This conclusion reinforces the growing opinion of the authors that the Heidke Skill Score has little to recommend it as a forecast verification statistic. A good verification statistic for subjective categorical forecasts should be easily optimized by the forecaster. This is not true of the Heidke Skill Score; thus comparison of different forecasters with this statistic is not valid if they were not equally informed on methods of maximizing the Skill Score. This is also not a valid statistic for comparing subjective categorical forecasts made in ignorance of the Bryan method with categorical forecasts derived objectively by means of the Bryan method from probability forecasts.

ACKNOWLEDGMENTS

The authors wish to thank Mr. Philip Williams for supplying the data used in this study and The Travelers Research Center, Inc., for lending the MDA computer program. The participation of the meteorologists who made the subjective analyses is greatly appreciated.

REFERENCES

1. G. W. Brier, "A Study of Quantitative Precipitation Forecasting in the TVA Basin," *Research Paper* No. 26, U.S. Weather Bureau, Nov. 1946, 40 pp.
2. G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, No. 1, Jan. 1950, pp. 1–3.
3. J. G. Bryan, "A Method for the Exact Determination of the Characteristic Equation and Latent Vectors of a Matrix with Applications to the Discriminant Function for More than Two Groups," Unpublished Ed. D. dissertation, Harvard University, 1950, 290 pp.
4. J. G. Bryan, "The Use of Probability Forecasts to Maximize Various Skill Scores," Unpublished manuscript, The Travelers Research Center, Inc., Hartford, Nov. 1962, 14 pp.
5. R. G. Miller, "Statistical Prediction by Discriminant Analysis," *Meteorological Monographs*, vol. 4, No. 25, American Meteorological Society, Boston, Oct. 1962, 54 pp.
6. C. R. Rao, *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, New York, 1952, 390 pp.
7. P. Williams, Jr., "Relationship of Precipitation to Vorticity and Vertical Motion at Salt Lake City, Utah," *Monthly Weather Review*, vol. 91, No. 5, May 1963, pp. 259–263.

[*Received July 17, 1964*]